TEXT-DRIVEN MOUTH ANIMATION FOR HUMAN COMPUTER INTERACTION WITH PERSONAL ASSISTANT

Yliess HATI *Francis ROUSSEAUX* *Clement DUHART*

Leonard de Vinci
Pole Universitaire, Research Center
Paris La Defense, 92916, France
yliess.hati@devinci.fr

URCA CReSTIC
Moulin de la Housse
Reims, 51100, France
francis.rousseaux@univ-reims.fr

MIT MediaLab
Responsive Environments Group
Cambridge, 02139, USA
duhart@mit.edu

ABSTRACT

Personal assistants are becoming more pervasive in our environments but still do not provide natural interactions. Their lack of realism in term of expressiveness and their lack of visual feedback can create frustrating experiences and make users lose patience. In this sense, we propose an end-to-end trainable neural architecture for text-driven 3D mouth animations. Previous works showed such architectures provide better realism and could open the door for integrated affective Human Computer Interface (HCI). Our study shows that such visual feedback improves users' comfort for 78% of the candidates significantly while slightly improving their time perception.

1. INTRODUCTION

Recent developments in the Artificial Intelligence (AI) community – more precisely in Deep Learning – re-enhanced affective computing with the promise of new communication layers between human and machine. This research area explores how computers can sense, analyze, generate, and express affect features as humans do. Sense and analysis of users' emotional states received much attention in the research community, especially for facial expression recognition [1], body gesture recognition [2], speech recognition, and natural language processing [3]. Usually, the identification of complex human affect expressions requires the use of multimodal frameworks or data fusion techniques [4], and the current state of the art allows the software to be responsive to the user's emotional states. However, affective Human Computer Interface (HCI) would also benefit from giving this kind of abilities to the computer. On the one hand, the computer should be able to generate an internal affective state in response to the user's interactions. On the other hand, it should be able to express it more naturally and realistically. Therefore, the user's biases regarding computer interaction could be reduced and would allow more credible communication loop-back between human and machine. Several contributions have studied the use of emotions during interactions with virtual agents, especially in negotiation [5, 6]. At our best knowledge, all of them used hard-coded fixed sets of expressions, whereas human's emotions is a continuous space. Our work aims to provide visual and acoustic expression abilities to computers. Figure 1 presents our workspace setup where users can interact with a personal assistant using our text-driven mouth framework. In this contribution, we propose an end-to-end architecture for voice synthesis with its associated mouth animations. This work is a preliminary step toward affect controls.

Figure 1: Personal Assistant based on our text-driven mouth animation system in a workspace setup.

2. RELATED WORK

Over the last decade, facial animation received considerable attention from industries and research communities for reducing their production cost and their fidelity. At the early days of this filed, facial animations were made by hand and required the expertise of professional animators. The 3D models were first rigged, weighted and then animated frame by frame. This process is time-consuming and varies with the animators. More recently, the use of performance capture allowed to semi-automatize facial animations. Facial movements are recorded with specific software and hardware by tracking markers on human actors in real-time. The collected data is then transferred to the 3D character model. The results are highly dependent on the actor morphology and appearance and require intensive cleaning [7, 8, 9]. Nowadays, thanks to the recent enhancements in Deep Learning, video and audio based approaches are gaining popularity.

The overall process requires the system to produce mouth ani-

https://doi.org/10.21785/icad2019.032

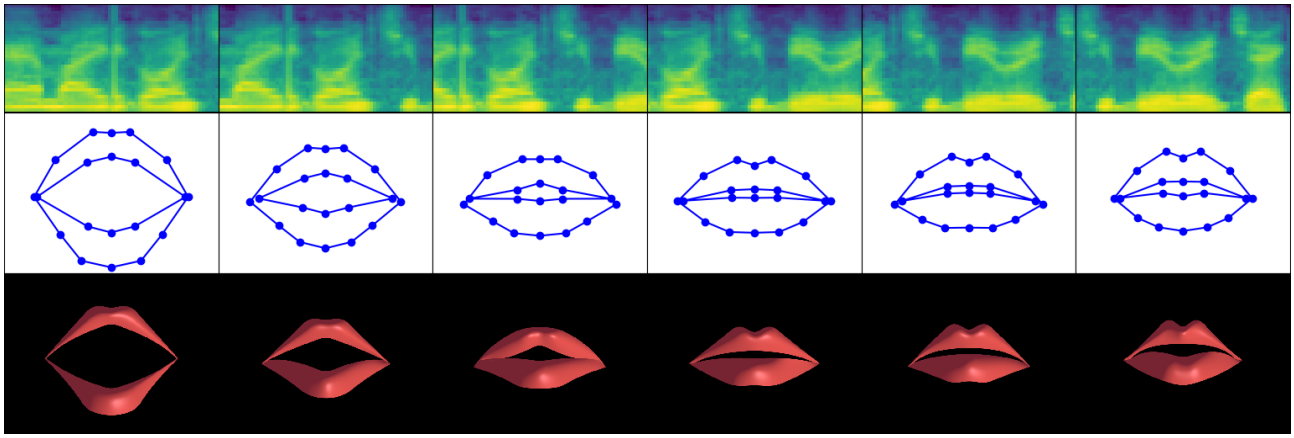

Figure 2: The first line represents the successive 40ms Mel-spectrograms computed from a testing video audio track with the corresponding inferred 3D mouth landmarks on the second line which have been used to generate the final rendering on the last line.

mations synced with the speech. One method is to define a dictionary of visual mouth forms called visemes. Then, the system needs to learn how to associate those visemes with the right phonemes and smooth the result using interpolation. Such approach discretizes the visual and acoustic spaces to bind them. Early works proposed the use of Hidden Markov Model [10, 11] whereas modern ones are using Deep Learning techniques [12, 13, 14, 15]. One limitation is the lack of expressiveness in the mouth articulations, limiting the realism of such generated animations. Moreover, audio and phonemes alignment is a complicated task and is not always feasible.

Nowadays, the Deep Learning community has been able to solve similar issues in translation. Instead of discretizing the input and output space to find bindings, the system learns how to map those spaces together directly. Hence, results are consistent in the output space. In the animation field, recent contributions present significant improvements using such approaches known as audio-driven or speech-driven facial animation. Such frameworks allow to synthesize facial movements and can include an emotional dimension [16, 17]. Unfortunately, the authors did not share their dataset, limiting reproducibility for future contributions.

Recently, ObamaNet from Rithesh and al. introduced the first approach for text-driven mouth animations [18]. Using text as input, their model can generate an audio waveform and its photo-realistic lip-sync frames. Their results are impressive and can lead to future approaches. However, this method is not suitable for other applications requiring the control of a 3D character model.

Our contribution includes:

- A methodology for the creation of text-driven mouth animation datasets from a video bank, including audio. First, the speaker's face is automatically detected and cropped from the video. Then, facial 3D landmarks are identified and projected into an invariant space. Finally, the mouth's key points are extracted and normalized, providing natural synchronicity with the audio.
- A new Deep Learning pipeline for text-driven mouth animation following best practices from state of the art. The speech synthesis module has been replaced by more recent contributions. We also introduce a new module to focus on 3D landmarks controls instead of 2D rendered frames.
- An experiment evaluating how our proposed generated mouth visuals can impact users' comprehension during a listening test and their realism during a blind test compared to landmarks extracted from real videos.

For research reproducibility, we will soon publish our generated dataset as well as our source code with public access.

3. ADMA DATASET

In this section, we describe the methodology used to generate our dataset ADMA-TED for Audio-Driven Mouth Animation (ADMA) TED based on the Lip Reading Sentence (LRS3-TED) dataset [19]. This dataset is composed of 400 hours of TED and TEDx English talks videos distributed over 4004 videos for training and validation, and 451 for testing. Each video varies from 1 to 6 seconds and is cropped on the speaker's face with a 224 by 224 pixels resolution. This dataset has been chosen for its diversity and quality, such as the faces are almost always visible continuously. The rejection rate is lower than 1% and concerns recordings with microphone glitches or where the face is not visible enough. For each video, we provide 20 normalized 3D mouth landmarks every 40 ms to ensure continuous face tracking during the sentence pronunciation, as illustrated in Figure 2.

3.1. Mouth Landmarks

Face Alignment Network (FAN), proposed in Bulat and al. [20], is a popular model for face landmark inference on pictures. It has been applied on the entire LRS3-TED dataset at 25 FPS rate, providing 68 3D face landmarks for each frame as illustrated Figure 3.

Only mouth landmarks are extracted, as illustrated in Figure 4. They are then projected into a head rotation invariant and normalized space. We defined the top of the nose and the chin as our y axis, both eye's landmarks as our x axis and their cross product as our z axis. These axes define the face referential. Using the transformation matrix TM in Eq. 1, the mouth's landmarks are projected into a new front-facing referential. These landmarks are normalized between $[0;1]$ with the center of the mouth located at 0.5 in each axis.

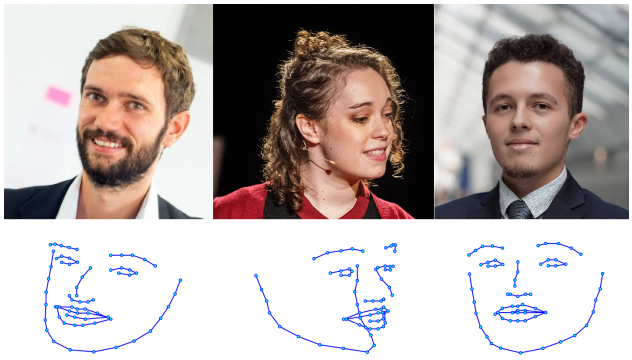

Figure 3: Examples of extracted landmarks in 3D coordinates using FAN algorithms on cropped faces.

$$TM(x,y,z) = \begin{bmatrix} x_0 & y_0 & z_0 & 0 \\ x_1 & y_1 & z_1 & 0 \\ x_2 & y_2 & z_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

Finally, as people possess different types of mouth, identity is removed by computing the median lips thickness over the entire dataset to reduce this bias. Depending on given thresholds, some videos are rejected for being out of the maximum and minimum range of the mouths' opening and closing.

3.2. Audio Processing

Each video's audio channel from LRS3-TED is sampled at 16kHz. Their Mel-spectrogram are computed using *32 frequency bands*, *128 hop length*, and *512 window size*. Each extracted mouth is associated with a centered *64 window size* of the Mel-spectrogram and filtered to remove high-frequency glitches. Each datum of the dataset contains a spectrogram with its associated mouth landmarks. The dataset is composed of 1,032,219 elements for training and 35,473 for testing.

4. END-TO-END NEURAL NETWORK ARCHITECTURE

Our end-to-end neural network for text-driven mouth animations is based on the architecture proposed by Rithesh and al. [18]. As state of the art has improved over the years, we have upgraded the Text-to-Speech (TTS) module. Our neural module for audio-driven mouth 3D landmarks regression is connected to the end of the processing pipeline, as illustrated in Figure 6. The 20 3D mouth landmarks are used in a 3D engine to compute mouth's vertices and normals frame by frame using 3D spline interpolation, as shown in Figure 5. The final rendering is achieved with the use of a Phong shader.

4.1. Mouth Landmark Regression

This section details the mouth landmarks regression module in charge of estimating mouth 3D landmarks coordinates conditioned on 40 ms Mel-spectrograms from speeches.

For this task, we considered two aspects: the mouth landmarks 3D spatial positions and their dynamics over time. Hence, our model contains three stages. The first stage learns frequency correlations in a compact feature maps representation over time. Then, the second stage determines the correlation dynamic between these step representations. Both stages use convolution layers with rectangle kernels such as time is considered as a spacial dimension instead of a sequence, allowing faster training and inference performances. Stride is preferred over pooling layers to keep as much information as possible. Finally, these compact features are used to train a Mean Squared Error (MSE) regression to estimate the 3D normalized landmarks positions.

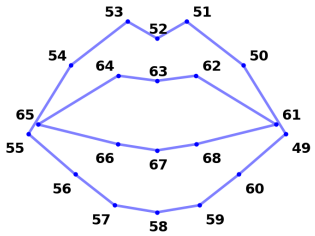

Figure 4: Illustration of the 20 points composing the mouth skeleton with their corresponding FAN annotations.

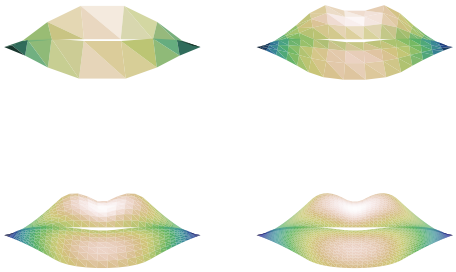

Figure 5: 3D Spline interpolation applied on the 20 mouth landmarks. Each image represent a different interpolation resolution. From left to right and top to bottom each 3D plot respectively correspond to interpolation factors of (8, 1), (16, 4), (32, 8) and (64, 16) with horizontal interpolation first and vertical interpolation second.

4.2. Text-Driven Mouth Skeleton

Rithesh and al. [18] proposed the first contribution of end-to-end neural architecture for text-driven lip-syncing. Their work inspired our proposed architecture. We upgraded their architecture with state of the art contributions and extended it from 2D space coordinates to 3D ones. We replaced the Char2Wav module by Tacotron2 [21] to convert text input into a Mel-spectrogram and WaveGlow [22] for phase reconstruction.

4.3. Neural Architecture Pipeline

Our proposal is an end-to-end neural architecture for text-driven mouth animations composed of different modules. Each one of

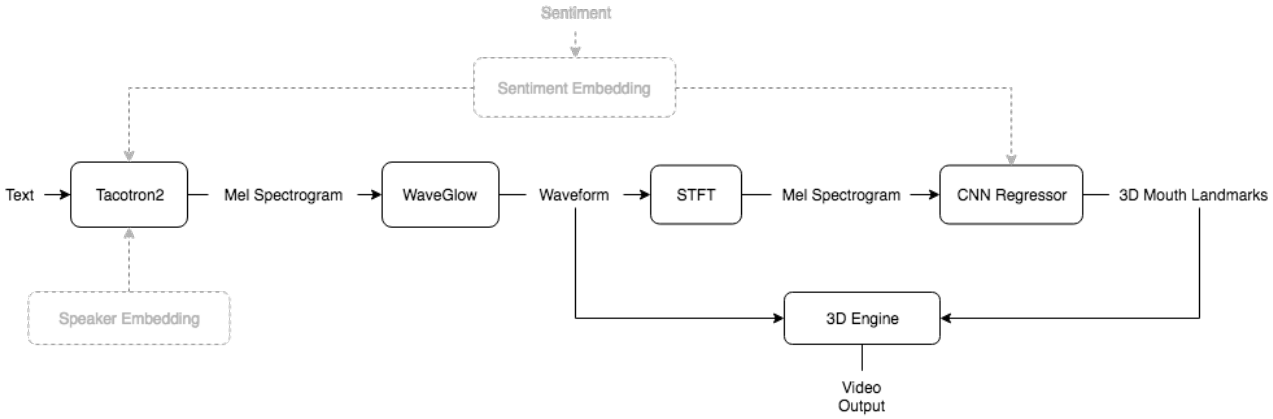

Figure 6: The overall architecture is composed of successive neural modules. The modules Taoctron2 and WaveGlow are in charge of the text to speech transformation. The following ADMANet module computes the 3D mouth landmarks animation based on the speech waveform. Finally, a rendering engine generates the final animation with its corresponding audio file.

those is a crucial step because of cumulative approximation errors throughout the entire model. During development, we tried different approaches for the pipeline. One of them was to connect our CNN regressor directly to the Tacotron2 output. However, the output resolution was not sufficient to allow our model to learn the mouth landmarks properly. The Mel-spectrogram use by our architecture is computed using an STFT on the waveform produced by WaveGlow, as illustrated in Figure 6, to allow control on its resolution.

Layer	Kernel	Stride	Outputs	Activation
Frequency Domain				
Conv2D	-	-	1x64x32	-
Conv2D	1x3	1x2	72x64x16	ReLU
Conv2D	1x3	1x2	108x64x8	ReLU
Conv2D	1x3	1x2	162x64x4	ReLU
Conv2D	1x3	1x2	243x64x2	ReLU
Conv2D	1x2	1x2	256x64x1	ReLU
Time Domain				
Conv2D	3x1	2x1	256x32x1	ReLU
Conv2D	3x1	2x1	256x16x1	ReLU
Conv2D	3x1	2x1	256x8x1	ReLU
Conv2D	3x1	2x1	256x4x1	ReLU
Conv2D	4x1	4x1	256x1x1	ReLU
Mouth Landmark Regression				
Dropout 0.5	-	-	256	-
FC	-	-	256	Sigmoid
FC	-	-	256	Sigmoid
FC	-	-	$20*3=60$	Linear

Table 1: ADMA-Net architecture is composed of three stages. First one is in charge of learning frequency correlations at each time step whereas the second stage learns their dynamics according to the incoming Mel-spectrogram. Finally, the last stage learns a regression between these frequency dynamic feature maps and the mouth landmark positions.

4.4. Training

Our model has been trained for 5 hours over 1000 epochs using a single *Nvidia 980M 8Go GPU*. Using the Adma optimizer with a $1e^{-3}$ *learning rate* and $(0.9, 0.999)$ *betas*, we achieved an error of $1e^{-3}$ on the *trainval* set and an error of $2e^{-3}$ on the *test* set using a simple MSE as our objective function.

5. EVALUATION

To asses the quality of our results, we conducted a blind user study. We wanted to evaluate the realism of our generated mouth animations and the impact of visual feedback on user's time perception and comprehension when listening to a potential personal assistant.

- The realism of our artificial mouth animations is evaluated by comparing them to mouths generated by a landmark tracking system applied to real videos. We considered two scenarios. In the first one, these two video categories have to be distinguished independently. In the second one, both classes have to be discriminated by pairs.
- The impact of visual feedback on user experience is evaluated through a listening and comprehension test. In this test, the user answers questions about recordings with or without mouth animations. The user must also estimate the speech's length which provides indications of its patience level.

5.1. Setup

The evaluations have been conducted in an open space environment on our experimental table presented in Figure 1. Candidates are confronted alternatively to recordings with and without the mouth animations over our different experiments presented as following.

The 24 candidates sampled for this experiment are randomly selected among English-speakers, including native ones, based on their exposure to 3D animations in the form of movies, videos games and modeling independently of their age and gender. We defined exposure as three categories: rare, casual, and daily. Finally, to avoid any bias in the results, professional animators and 3D modelers are excluded from the candidates.

	Age		Exposure to 3D animations		
	18-25	*26-59*	*Rare*	*Casual*	*Daily*
Man	8	4	2	2	8
Woman	4	8	4	6	2
Total	12	12	6	8	10

Table 2: Study candidate composition.

5.2. Realism Evaluation

This evaluation estimates how our system can fool candidates. For a given audio recording, candidates need to identify and classify artificial mouth animations from tracked ones. In this sense, we confronted them to two experiments.

- Candidates are exposed to ten independent mouth animations from 3 to 6 seconds in random order and have to assign them to the appropriate category. Each candidate took 2 to 5 minutes to complete the task. In Figure 7, the median success rate among all profiles is around 50% with a standard deviation of 14.2%. Results show that candidates are not able to correctly identify the videos as being part of one of the two classes.
- To extend our experiment further and assess the robustness of the results, users are then exposed to five pairs of 3 to 6 seconds of videos from each category and have to distinguish them side by side. This task took 3 to 7 minutes to complete for each candidate. In Figure 8, candidates are still not able to discriminate the two mouth animation generation methods.

According to Figure 7, and Figure 8, gender and age do not seem to impact candidates' ability to solve the identification and classification tasks. Contrary, the exposure criteria tend to help the users. Therefore, as illustrated in Figure 9, both experiments confirm that our model can produce mouth animations realistic enough to fool human candidates.

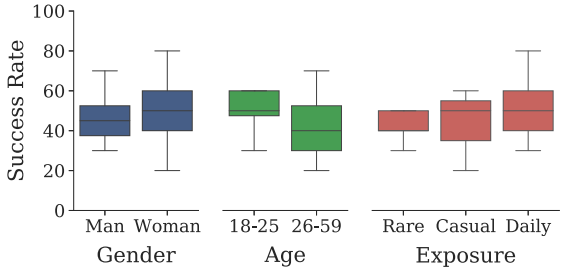

Figure 7: Success rate during blind independent identification test of artificially generated and human tracked mouth animations. Results are displayed per independent discrimination criterion: gender (in blue), age (in green) and exposure to 3D animations (in red). Each box-plot describes the maximum, first quartile, median, third quartile and minimum of the candidates' success rate.

5.3. Comprehension Evaluation

Visual feedback is capable of affecting the human's comprehension capacity. To evaluate the impact of our mouth animations on comprehension, we realized two experiments. Sixteen audio recordings – from 3 to 18 seconds randomly placed in 20 seconds tracks – with or without visual feedback, were shown to the user who needed to answer a comprehension question with four choices. In the first one, the user is asked to answer a question concerning the content of the speech. In the second one, an answer is shown to the user who needs to find the correct item among four possible questions, allowing to perform cross-validation on the visual feedback impact. We expect our mouth animations to alter candidates' cognition.

Results of this evaluation do not show any significant difference when visual feedback is provided or not, thus for any candidate profile.

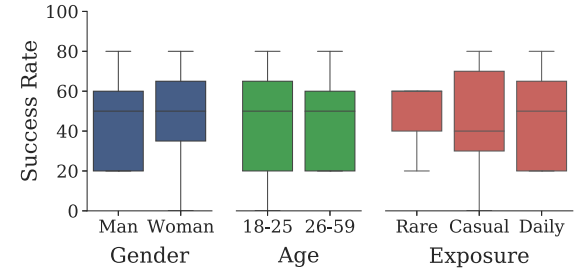

Figure 8: Success rate during blind side by side discrimination test of artificially generated and human tracked mouth animations. Results are displayed per independent discrimination criterion: gender (in blue), age (in green) and exposure to 3D animations (in red). Each box-plot describes the maximum, first quartile, median, third quartile and minimum of the candidates' success rate.

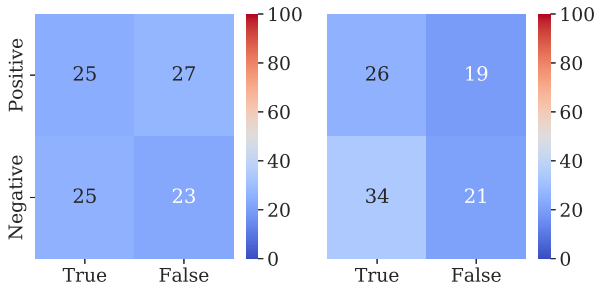

Figure 9: Confusion matrices during blind tests for respectively independent identification (on the left) and side by side discrimination (on the right) of artificially generated and human tracked mouth animations. Each value have been transformed into percentages.

5.4. Time Perception

Time perception can be an indicator of a user's patience during listening tests. Our generated mouth animations are expected to increase people's ability to measure this component by providing visual feedback. To this end, candidates were asked to estimate the duration of recordings during the comprehension tests. The audio recordings from the comprehension evaluation task are independently displayed one by one with or without visual feedback. Both comprehension and evaluation tasks took 7 to 10 min to complete in total.

In Figure 11, results show a small difference between candidates estimation errors with and without visual feedback. Compared to audio only, visual feedback slightly improves candidates ability to estimate the recordings' duration. For more details, we provide a histogram for each question Figure 10. These histograms show that visual feedback tends to increase candidates accuracy for the duration estimation task. Answers are less sparse and more centered around the right one than with audio only.

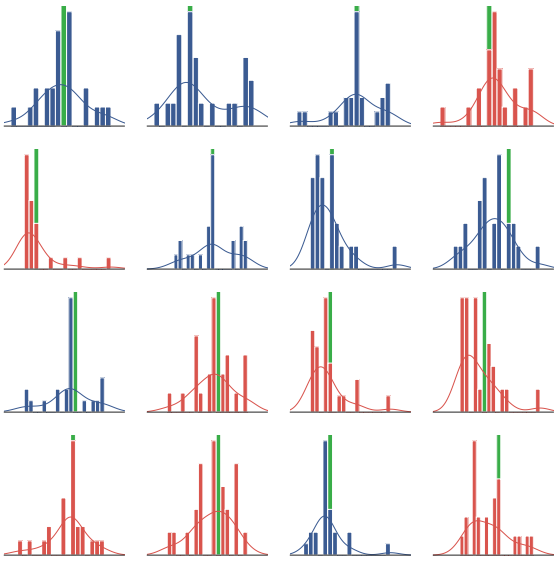

Figure 10: Histograms of speech duration estimation for audio only (in blue) versus audio with visual feedback (in red). Each histogram corresponds to a different speech. Speech duration varies from 3 to 18s. Original answer for the speech duration (in green) and histogram tendency curve can be observed on the graph.

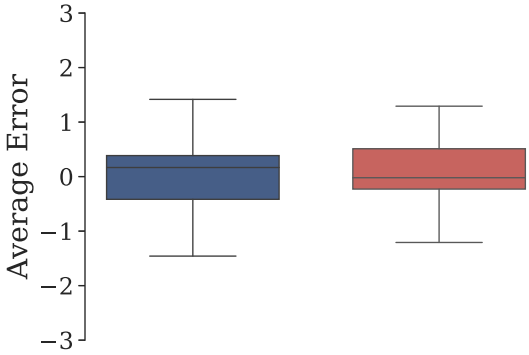

Figure 11: Average error of the speech duration estimation by the candidates during comprehension test between audio only (in blue) and audio with visual feedback (in red). Each box-plot describes the maximum, first quartile, median, third quartile and minimum of the candidates' duration estimation difference from the right answer.

5.5. Interactivity Comfort

The ultimate perspective of this work is interactivity comfort with personal assistants. This contribution focuses on the interest of visual feedback during conversations between users and personal assistants. Hence, it has been challenging to evaluate how visual feedback improves interactions independently of the personal assistant's limitations. In this sense, we let candidates rate their use of a Google Assistant with and without our ADMA interface. The average rating shows 78% of the candidates estimated that comfort is significantly improved with visual feedback, independently of their gender, age or exposure to 3D animations.

5.6. Synthesis

Our proposed evaluation shows that the generated mouth animations combined to speech synthesis are realistic enough to fool, at list partially, human candidates. If initial expectations considered visual feedback as an approach to more natural HCIs, our study could not allow concluding quantitatively either qualitatively such assumptions. However, we observed that the candidates have a better estimation of the recording's length in the presence of mouth animations. Hence, we can assume that visual feedback improves time perception and comfort independently of the gender, age, and exposure to 3D animations.

6. CONCLUSION

In this work, we present an end-to-end Text-driven Mouth Animation model. Contributions in generative animations traditionally benefit the entertainment sector for video games and animated movies. We explore such technology to the use of virtual avatars for personal assistants. Visual feedback often enables more natural ways to interact, especially when combined with audio. Our focus in this work is the generation of text-driven mouth animations for such personal assistants to improve the user experience.

To this end, we proposed a methodology for the creation of datasets dedicated to text-driven 3D mouth animations. This approach does not require the use of performance capture and can easily be extended to future applications. We also developed an end-to-end neural network model combining state of the art in speech synthesis and a custom neural regressor inspired by previous work trained on our dataset. This model allows the creation of 3D mouth animations with its associated speech audio conditioned on unique text input. We also provided an evaluation set to estimate the realism of the generated mouth animations and their value in terms of improvement for HCI. We conclude that the level of realism is good enough to fool human candidates, whereas it does not help in solving cognition task of comprehension. The interest seems limited to improving user comfort.

6.1. Possible Applications

We consider some application domains in which our model and study can be relevant for future developments.

- Video Games and animated movies could use our text-driven mouth animations to animate their characters or as placeholders for preview animations proving a better sense of the final rendering. Compared to performance capture, our system does not require the use of sophisticated software and/or hardware

that some could not afford. Our proposal also provides consistent results and does not depend on the artist's animation skills.

- Art museums and exhibitions could use text-driven mouth animations to renew how visitors can experience art galleries. Text-driven portraits or even text-driven description cards could bring interactivity to traditional audio guides. Such technology could be extended with other visual mouth renderings instead, the one presented in this work.
- People suffering from Autism spectrum disorder (ASD) could also benefit from such technology. Controlled robotic interaction has proven to be less frustrating among these individuals and is used as a new form of therapy. Children with ASD are better at integrating audiovisual feedback compared to real people during social interactions. As our work focus on the importance of visual feedback for the creation of less frustrating natural interactions, our model could benefit those robots and enhance their ability.

6.2. Limitations

Major limitations observed in our work are directly inherited from deep learning sensitivity to the data it is trained on and its heavy computation costs.

- The visual realism fidelity of our mouth animations seems affected by the lack of horizontal movement. Our model does not seem to capture the horizontal mouth movements as good as the vertical ones. To tackle this issue, we plan to retrain our regressor network using different regularization techniques.
- The computation cost of our pipeline is limited by the requirement of a powerful GPU. It cannot be suitable in some applications requiring limited physical space, power consumption, and network constraints. Hence, two aspects of the pipeline could be improved. On the one hand, WaveGlow, used in the TTS module, is a bottleneck for being able to integrate such project on Sytem On a Chip (SOC) devices and could be optimized by different approaches such as batch inference. On the other hand, our visual rendering engine is limited by the mouth's 3D vertices and normals computation occurring on each frame and could be optimized by re-targeting the 3D landmarks on a 3D mouth model instead.
- The unique voice used in our Speech Synthesizer can induce biases and is not suitable for applications requiring different gender, race or age characteristics or even neutrality. We plan to address this issue by retaining and conditioning each module of the TTS on a speaker latent variable using a multi-speaker dataset.

6.3. Future Work

One extension of this work is the introduction of an emotional and prosody context to the entire system. The emotional context of the speaker influences lips movement, speech tones, and styles. We think avatars and personal assistants should be able to express this kind of features and would result in better interactions with humans. Previous works show that a sentiment and prosody embedding or latent context variable can leverage such controls over TTS and Facial Animations independently [17, 23]. We want to extend our work using this approach of a learned latent variable as a way to influence the whole system in its integrity.

7. REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018. [Online]. Available: http://arxiv.org/abs/1804.08348

[2] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[3] F. Weninger, M. WÃűllmer, and B. Schuller, *Emotion Recognition in Naturalistic Speech and Language - A Survey*. John Wiley & Sons, Ltd, 2015, ch. 10, pp. 237–267.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98 – 125, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517300738

[5] C. M. de Melo, P. Carnevale, and J. Gratch, "The effect of expression of anger and happiness in computer agents on negotiations with humans," in *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, ser. AAMAS '11. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 937–944. [Online]. Available: http://dl.acm.org/citation.cfm?id=2034396.2034402

[6] ——, "The effect of virtual agents' emotion displays and appraisals on people's decision making in negotiation," in *Intelligent Virtual Agents*, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 53–66.

[7] W.-C. Ma, M. Lamarre, E. Danvoye, C. Ma, M. Ko, J. von der Pahlen, and C. A. Wilson, "Semantically-aware blendshape rigs from facial performance measurements," in *SIGGRAPH ASIA 2016 Technical Briefs*, ser. SA '16. New York, NY, USA: ACM, 2016, pp. 3:1–3:4. [Online]. Available: http://doi.acm.org/10.1145/3005358.3005378

[8] A. Smith, M. Sanders, C. A. Wilson, S. Pohle, W.-C. Ma, C. Ma, X.-C. Wu, Y. Chen, E. Danvoye, J. Jimenez, and S. Patel, "Emotion challenge: building a new photoreal facial performance pipeline for games," 07 2017, pp. 1–2.

[9] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: ACM, 2011, pp. 77:1–77:10. [Online]. Available: http://doi.acm.org/10.1145/1964921.1964972

[10] D. Cosker, D. Marshall, P. L. Rosin, and Y. Hicks, "Speech driven facial animation using a hidden markov coarticulation model," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1 - Volume 01*, ser. ICPR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 128–131. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.851

[11] L. Xie and Z.-Q. Liu, "Speech animation using coupled hidden markov models," vol. 1, 01 2006, pp. 1128–1131.

[12] Y. Zhou, S. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *CoRR*, vol. abs/1805.09488, 2018. [Online]. Available: http://arxiv.org/abs/1805.09488

[13] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. M atthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 93:1–93:11, Jul 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073699

[14] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1283–1302, Oct. 2005. [Online]. Available: http://doi.acm.org/10.1145/1095878.1095881

[15] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 127:1–127:11, Jul 2016. [Online]. Available: http://doi.acm.org/10.1145/2897824.2925984

[16] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3d facial animation from raw waveforms of speech," *CoRR*, vol. abs/1710.00920, 2017. [Online]. Available: http://arxiv.org/abs/1710.00920

[17] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 94:1–94:12, Jul 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073658

[18] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *CoRR*, vol. abs/1801.01442, 2018. [Online]. Available: http://arxiv.org/abs/1801.01442

[19] T. Afouras, J. Son Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv e-prints*, p. arXiv:1809.00496, Sept. 2018.

[20] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," *arXiv preprint arXiv:1712.02765*, 2017.

[21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[22] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," *arXiv e-prints*, p. arXiv:1811.00002, Oct. 2018.

[23] Y. Lee, A. Rabiee, and S. Lee, "Emotional end-to-end neural speech synthesizer," *CoRR*, vol. abs/1711.05447, 2017. [Online]. Available: http://arxiv.org/abs/1711.05447